

DOCUMENT RESUME

ED 344 902

TM 018 207

AUTHOR Mazzeo, John; And Others
TITLE Comparability of Computer and Paper-and-Pencil Scores
for Two CLEP General Examinations. College Board
Report No. 91-5.
INSTITUTION College Entrance Examination Board, New York, N.Y.
SPONS AGENCY Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-14
PUB DATE 91
NOTE 24p.
AVAILABLE FROM College Board Publications, Box 886, New York, NY
10101-0886 (\$12.00).
PUB TYPE Reports - Research/Technical (143) --
Tests/Evaluation Instruments (160)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *College Entrance Examinations; College Students;
Comparative Testing; *Computer Assisted Testing;
English; Higher Education; *Mathematics Tests;
*Scores; Test Construction; Test Format; Writing
(Composition); *Writing Tests
IDENTIFIERS *College Level Examination Program; English
Composition Test; *Paper and Pencil Tests

ABSTRACT

Two studies investigated the comparability of scores from paper-and-pencil and computer-administered versions of the College-Level Examination Program (CLEP) General Examinations in mathematics and English composition. The first study used a prototype computer-administered version on each examination for 94 students for mathematics and 116 for English. Based on these results and participant feedback, modifications were made to the prototype versions, and a second study was conducted using the modified tests for 96 mathematics students and 115 in English composition. Both studies used a single group counterbalanced equating design. Mathematics data were collected at Southwest Texas State University (San Marcos) and English results were collected at Utah State University (Logan). Study 1 results suggest that there were changes in average scores as a function of test delivery. Study 2 results indicate that modifications to the computer versions eliminated the mode-of-administration effects for the English composition examination, but not for the mathematics examination. Results of both studies underscore the need to determine empirically the equivalence of computer and paper versions of an examination. Twenty-four data tables, 8 figures, 10 references, and 2 appendices containing the questionnaires for student background information are included. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED344902

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL
IN OTHER THAN PAPER COPY HAS BEEN
GRANTED BY

ROBERT G. CAMERON

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

College Board
Report



No. 91-5

Comparability of Computer and Paper-and-Pencil Scores for Two CLEP[®] General Examinations

John Mazzeo
Barry Druesne
Paul C. Raffeld
Keith T. Checketts
Alan Muhlstein

BEST COPY AVAILABLE

**Comparability
of Computer
and Paper-and-Pencil
Scores for Two CLEP®
General Examinations**

**John Mazzeo
Barry Druesne
Paul C. Raffeld
Keith T. Checketts
Alan Muhlstein**

**College Board Report No. 91-5
ETS RR No. 92-14**

College Entrance Examination Board, New York, 1991

John Mazzeo is Analysis Director, NAEP Trial State Assessment, Division of Statistics and Psychometrics, at ETS.

Barry Druesne is a Program Administrator, College Board Testing Programs, at ETS.

Paul C. Raffeld is an Associate Professor, Department of Psychology, at Southwest Texas State University.

Keith T. Checketts is a Professor, Department of Psychology, at Utah State University.

Alan Muhlstein is a Research Associate, Department of Psychology, at Utah State University.

Acknowledgments

The work described herein was funded by the College Board/Educational Testing Service (ETS) Joint Staff Research and Development Committee and by the College Board's College-Level Examination Program (CLEP). The authors would like to thank Bernadette McIntosh, Claudia Whitehurst, and Robert Foy of ETS for their help in data file preparation and the production of figures; Doris Johnson of ETS for her help in editing the manuscript; Martha Stocking and William C. Ward of ETS for their reviews and comments on an earlier draft of the report; and Anne Mavor at the National Academy of Sciences for her help in the planning and execution of the studies. In addition, the authors would like to extend special thanks to Orlando Toro of the College Board for his leadership, support, and encouragement throughout a somewhat long and sometimes frustrating project.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,800 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is \$12.

Copyright © 1991 by College Entrance Examination Board. All rights reserved.

College Board, College-Level Examination Program, CLEP, and the acorn logo are registered trademarks of the College Entrance Examination Board.

Printed in the United States of America.

CONTENTS

Abstract	1
Introduction	1
Background	1
Description of Examinations	2
Mathematics Examination	2
English Composition Examination	2
Method	3
Results of Study 1	4
Mathematics Examination	4
English Composition Examination	8
Summary of Study 1 Results	9
Modifications to the Computer Versions for Study 2	10
Results of Study 2	10
Mathematics Examination	10
English Composition Examination	13
Summary of Study 2 Results	15
Discussion and Conclusions	15
References	16
Appendix A. Computer Familiarity Questionnaire (Study 1)	17
Appendix B. Computer Familiarity Questionnaire (Study 2)	18

Figures

1. Mathematics total examination scaled scores (Study 1)	4
2. Mathematics Section 1 scaled scores (Study 1)	5
3. Mathematics Section 2 scaled scores (Study 1)	5
4. English total examination scaled scores (Study 1)	9
5. Mathematics total examination scaled scores (Study 2)	11
6. Mathematics Section 1 scaled scores (Study 2)	11
7. Mathematics Section 2 scaled scores (Study 2)	12
8. English total examination scaled scores (Study 2)	14

Tables

1. Mathematics Examination Results, Both Versions (Study 1)	4
2. Mathematics Examination Results, Both Versions, by Order of Administration (Study 1)	5
3. Mathematics Examination Results by Section for Both Versions and Each Group (Study 1)	5
4. Score Correlations and Reliability Estimates for Mathematics Examination Scores (Study 1) ...	6
5. CFQ ₁ Scores of Mathematics Examinees (Study 1)	6
6. Correlations between CFQ ₁ Scores and Mathematics Examination Scores (Study 1)	7
7. Estimates of Coefficients and Squared Multiple Correlations for Models I and II for Section 1 and Section 2 Mathematics Examination Scores (Study 1)	7
8. Estimates of Regression Coefficients and Squared Multiple Correlations for Models III and IV for Section 1 and Section 2 Mathematics Examination Scores (Study 1)	8
9. English Composition Examination Results, Both Versions (Study 1)	8
10. English Composition Examination Results, Both Versions, by Order of Administration (Study 1)	9
11. Mathematics Examination Results, Both Versions (Study 2)	10
12. Mathematics Examination Results, Both Versions, by Order of Administration (Study 2)	11
13. Mathematics Examination Results by Section for Both Versions and Each Group (Study 2)	11
14. Score Correlations and Reliability Estimates for Mathematics Examination Scores (Study 2) ...	12
15. CFQ ₂ Scores of Mathematics Examinees (Study 2)	13
16. Correlations between CFQ ₂ Scores and Mathematics Examination Scores (Study 2)	13
17. Estimates of Coefficients and Squared Multiple Correlations for Models I and II for Section 1 and Section 2 Mathematics Examination Scores (Study 2)	13
18. Estimates of Regression Coefficients and Squared Multiple Correlations for Models III and IV for Section 1 and Section 2 Mathematics Examination Scores (Study 2)	13
19. English Composition Examination Results, Both Versions (Study 2)	14
20. English Composition Examination Results, Both Versions, by Order of Administration (Study 2)	14
21. CFQ ₂ Scores of English Composition Examinees (Study 2)	15
22. Correlations between CFQ ₂ Scores and English Composition Examination Scores (Study 2) ...	15
23. Estimates of Coefficients and Squared Multiple Correlations for Models I and II for English Composition Examination Scores (Study 2)	15
24. Estimates of Regression Coefficients and Squared Multiple Correlations for Models III and IV for English Composition Examination Scores (Study 2)	15

ABSTRACT

This report describes two studies that investigated the comparability of scores from paper-and-pencil and computer-administered versions of the College-Level Examination Program (CLEP) General Examinations in Mathematics and English Composition. The first study used a prototype computer-administered version of each examination. Based on the results of the first study and feedback from the study participants, several modifications were made to these prototype versions. A second study was then conducted using the modified computer versions. Both studies used a single-group counterbalanced equating design. Data for the Mathematics Examination were collected at Southwest Texas State University, and data for the English Composition Examination were collected at Utah State University. The results of Study 1 suggest that, despite efforts to design computer versions of the CLEP Mathematics and English Composition General Examinations that were administratively similar to the paper-and-pencil examinations (i.e., allowed item review and answer changing and were comparably timed), mode-of-administration effects (i.e., changes in average scores as a function of the mode of test delivery) were found. The results of Study 2 suggest that the modifications made to the computer versions eliminated the mode-of-administration effects for the English Composition Examination but not for the Mathematics Examination. The results of both studies underscore the need to determine empirically (rather than to just assume) the equivalence of computer and paper versions of an examination.

INTRODUCTION

The computer administration (both conventional and adaptive) of existing educational tests has been recognized for some time as an attractive alternative to their typical paper-and-pencil administration (Bunderson, Inouye, and Olsen 1989; Ward 1984). The simplest variant of a computerized test might be called a *linear computerized test* (LCT). In an LCT, a previously existing standardized test (typically a multiple-choice test) is adapted to be administered by computer. The entire set of items in the paper version of the test is administered to all examinees. Items are presented individually (or in small groups) on the computer screen in the same order as they appear in the paper version of the test. Some of the advantages of LCTs typically cited are increased test security; reduced costs of test production, administration, and scoring; and the possibility of immediate feedback of results.

As Bunderson et al. (1989) point out, LCTs are intended to serve the same purposes as their paper-and-pencil counterparts. Therefore, a key issue is the comparability of scores obtained on computer and paper versions of the same tests. Can the scores from the two versions be used interchangeably to make academic decisions?

This concern about comparability is explicitly acknowledged in the American Psychological Association

(APA) *Guidelines for Computer-Based Tests and Interpretations* (1986). Guideline 16 states:

When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (p. 14)

This report describes two studies that investigated the comparability of scores from paper-and-pencil and computer-administered versions of the CLEP General Examinations in Mathematics and English Composition. The first study used a prototype computer-administered version of each examination. Based on the results of the first study and feedback from the study participants, several modifications were made to these prototype versions, and a second study was conducted using the modified computer versions.

BACKGROUND

Beginning in 1987, the College Board and the Educational Testing Service developed a prototype linear computer version of each of the five CLEP General Examinations (Mathematics, English Composition, Humanities, Social Sciences and History, and Natural Sciences). Each examination is 90 minutes long and covers material usually taught in the first two years of college (often referred to as the general education or liberal arts requirement). Total formula scores on each examination (number of correct responses minus a fraction of the number of incorrect responses) are linearly converted to a score reporting scale of 200 to 800. For those examinations that report subscores, a similar procedure is used; however, the subscore reporting scale runs from 20 to 80. The results of CLEP examinations are most often used by college students to obtain credit for, or exemption from, college courses. Some colleges grant credit for or exemption from particular courses that cover material similar to that contained in the examinations. Others grant general credit toward the fulfillment of a liberal arts or distribution requirement in a particular subject area.

Computer versions of the examinations were created from entire 90-minute paper-and-pencil versions that had been previously equated to the CLEP score reporting scale. They were developed to run on IBM-compatible equipment with two floppy disk drives, 640K of RAM memory, a standard keyboard, and an EGA monitor. The computer versions were designed to be as similar as possible in administration format to the paper versions of the examinations. Most items were presented individually, in their entirety, on a single computer screen. However, a number of items using figures, tables, or graphs and the items based on reading passages could not be displayed on a single screen. For

these items, the figure, table, graph, or reading passage was presented on an initial screen and the question relating to this material was presented on a subsequent screen. Examinees were free to toggle back and forth between screens as often as required. The time limits for the computer and paper versions of the examinations were identical. Within each section of the computer versions of the examinations (most CLEP examinations consist of two separately timed sections), students were given the opportunity to skip items and return to them later. In addition, within each section, students were permitted to change their answers.

Before introducing the computer versions, it was necessary to determine whether scores from the computer and paper versions of the examinations could be used interchangeably. From the point of view of administrative ease and cost considerations, it was hoped that computer administration would have no effect and that the existing (i.e., paper-and-pencil) conversions of formula scores to the CLEP score reporting scales could be used with the prototype computer versions. However, a review of the literature on the equivalence of scores from automated and conventional tests (Mazzeo and Harvey 1988) suggested that it would be unwise to assume comparability of scores without empirical evidence, and the previously referenced APA guidelines clearly suggested that comparability should be empirically established. If a mode-of-administration effect was observed (i.e., a change in average examination scores as a function of the mode of test delivery), separate equatings of scores from the computer versions to the CLEP scales would have to be obtained before using these versions operationally.

As a first step toward examining the comparability of scores from computer and paper versions of the CLEP General Examinations, the studies of the Mathematics and English Composition Examinations described in this report were conducted. These examinations were chosen for two reasons. First, they are taken by the greatest number of candidates among the General Examinations, so it was desirable to make computer versions available as soon as possible. Second, because they consist of relatively simple multiple-choice items, it was hypothesized that the likelihood of mode-of-administration effects for the examinations would be small. If such effects were not in evidence, consideration could be given to operational use of the computer versions.

During the fall of 1987 an attempt was made to conduct a study of the comparability of the computer and paper versions of the Mathematics and English Composition Examinations using a national sample of candidates from institutions awarding CLEP credit. However, it was difficult to secure the degree of participation necessary to execute the study. As a result, study plans were modified and, in the summer of 1988, Southwest Texas State University (SWTSU) and Utah State University (USU) agreed to a joint research project in an effort to obtain preliminary data.

Since the fall of 1986, SWTSU has conducted a General Studies Examination Program that requires all students to pass college-level examinations in writing and mathematics before completing more than 15 hours of upper-division work. The examination chosen for the mathematics requirement was the CLEP General Examination in Mathematics.

Utah State University began awarding credit for the CLEP General Examinations in the early 1970s. The examinations are not required, but they are advertised and offered to all students who wish to attempt them, including freshmen entering directly from high school. USU offers credit in all CLEP General Examinations except Mathematics. Students who pass the CLEP English Composition Examination are given credit for six quarter hours of English and have met the freshman English writing requirement.

DESCRIPTION OF EXAMINATIONS

Mathematics Examination

The CLEP General Examination in Mathematics is intended to cover material generally taught in a college course for nonmathematics majors and majors in other fields not requiring a knowledge of advanced mathematics. The test contains 90 four-option multiple-choice items that are administered to students in two separately timed sections. The first section (Basic Skills and Concepts) contains 40 items covering arithmetic, algebra, geometry, and data interpretation. Students are given 30 minutes to complete this section. The second section (Content) contains 50 items covering sets, logic, the real number system, functions, and probability. Students are given 60 minutes to complete this section. Each section is formula scored, with examinees receiving one point for each correct answer, no points for omitted items, and losing one-third of a point for each incorrect answer. The total formula score on the examination is a weighted sum of the scores for the two sections, with the second section counting twice as much.

The total examination scores are reported on a scale of 200 to 800. In addition to the total examination scores, separate subscores are reported on a 20 to 80 scale for each section. A 60-minute version of the examination was scaled in 1972 to have a mean of 500 (50 for the subscores) and a standard deviation of 100 (10 for the subscores) for a reference population of college sophomores. The subsequent 90-minute version (introduced in 1978) was equated to previous forms and referenced to the same scales (College Board 1978).

English Composition Examination

The General Examination in English Composition is intended to measure the skills required in most first-year college composition courses. Two versions are offered: an all-multiple-choice version (consisting of 95 five-option items)

and a version consisting of 45 multiple-choice items and an essay. Only the all-multiple-choice version has been adapted for computer delivery.

The first section of the examination contains 55 items that focus on logical, structural, and grammatical relationships *within sentences*. Students are given 45 minutes to complete this section. The second section contains 40 items that focus on an analysis of the elements of language *within larger prose passages*. Many of the items in the second section are based on short prose passages. Students are given 45 minutes to complete this section. The English Composition Examination total scores are also reported on a scale of 200 to 800. However, no additional subscores are reported.

A 60-minute version of the examination was scaled in 1963 to have a mean of 500 and a standard deviation of 100 for a reference population of college sophomores. Content specifications were revised in 1986, and the all-multiple-choice 90-minute version was scaled to the 60-minute version using an anchor-test design (College Board 1986).

METHOD

Both SWTSU and USU participated in two separate rounds of data collection (referred to below as Study 1 and Study 2). The design of the two studies was identical. Therefore, a single section is included to describe the procedure used for both rounds of data collection. However, results for Studies 1 and 2 are described in separate sections.

At both institutions the limited number of available computers and the time frame for obtaining results restricted the number of students who could be tested. Because the number of students tested was expected to be small at both institutions, the study was conducted using a single-group counterbalanced equating design (design 2 in Angoff 1984).¹ In general, for a given sample size, greater precision is obtained from this design than from random groups or anchor-test designs.

Each student participating in the studies took both a 90-minute computer version and a 90-minute paper version of the examination. As mentioned above, the computer versions were created from existing paper-and-pencil forms that had been calibrated to the CLEP 200–800 score reporting scale. The paper versions used in the study were also previously existing (and calibrated) 90-minute forms. At SWTSU the computer and paper-and-pencil versions had no items in common. At USU they had 29 items in common.

Students were randomly assigned to one of two groups. One group (subsequently referred to as the computer-first group) took the computer version of the ex-

amination before taking the paper version. The second group (subsequently referred to as the paper-first group) took the paper version before taking the computer version. Upon completing the examinations, students from both groups were asked to fill out a computer familiarity questionnaire (CFQ) based on a similar questionnaire developed and used by Lee (1986). Slightly different versions of the questionnaire were used for Studies 1 and 2. The questionnaires for Study 1 and Study 2 are shown in Appendixes A and B, respectively. The Study 1 version consisted of six items, five of which came from the questionnaire developed by Lee. The sixth question was added by the current authors. The Study 2 version consisted of slightly modified versions of the Study 1 questions plus two additional questions developed by the current authors.

Students volunteering to participate at SWTSU were given the \$35 CLEP Mathematics Examination free of charge, and they received the highest of the two resulting scores (using the already existing paper-and-pencil formula score to scale conversions for the computer version). Apparently, this was a successful incentive, since 94 of 120 students who were tested as part of the General Studies Examination Program participated in Study 1. In Study 2, 96 of 110 students in the General Studies Examination Program participated.

At USU all entering students were sent information regarding credit by examination and were given the opportunity to take one or more of these examinations during a pre-college orientation visit to the campus. Students who elected to take the CLEP English Composition Examination were then given the chance to take both versions of the test at half the usual fee and received the highest of the two resulting scores.

At SWTSU, the paper version of the examination was given to all 120 students in Study 1 over a three-day period in the fall of 1988 and to all 110 students (Study 2) over a three-day period in the spring of 1989. The computer administrations began four days before the paper-and-pencil administrations and extended four days after. Each student was assigned a two-hour block of time for computer testing. Students were asked to complete a questionnaire about computer familiarity and were then allowed to work through sample questions on the computer before starting the timed sections of the examination.

At USU, the students participating in Study 1 were allowed to register for and take the examinations during the summer orientation period, which began in July 1988, and testing continued through the fall quarter. For Study 2, students were allowed to register in July 1989, and testing continued through February 1990. Each student who agreed to participate was required to complete both versions of the examination on the same day. Students were randomly assigned to the computer-first or paper-first groups, and the results of the examinations were not revealed until both versions had been completed. Each student also completed a computer familiarity questionnaire.

1. Initial plans included doing separate analyses by gender and by racial/ethnic subgroup membership. However, because of the relatively small size of the studies and a substantial degree of missing demographic data (particularly for gender), insufficient sample sizes were available for separate subgroup analyses.

RESULTS OF STUDY 1

Mathematics Examination

A total of 94 students (47 in the computer-first group and 47 in the paper-first group) agreed to participate in Study 1. The scaled score mean and standard deviation for the computer and paper versions of the examination are given in Table 1. Since all students took both versions and equal numbers of students took the computer and paper versions first, one might expect that the means and standard deviations for the two versions would be roughly the same if no mode-of-administration effects were present.² However, the mean of the computer-version scores was nine points lower than the mean of the paper-version scores. Although this difference is not large (.13 paper-and-pencil test standard-deviation units), it is statistically significant ($t = 2.09$, $df = 93$, $p < .05$).

Table 2 gives scaled score means, standard deviations, and the mean difference between the paper and computer versions of the Mathematics Examination separately for the computer-first and paper-first groups. For the computer-first group, the mean paper-version score was 23 points higher than the mean computer-version score. This difference is statistically significant ($t = 4.08$, $df = 46$, $p < .01$). For the paper-first group, the mean paper-version score was five points lower than the mean computer-version score. This difference is not statistically significant ($t = -.84$, $df = 46$). This information is also displayed in Figure 1, which provides a plot of the mean scores given in Table 2 by testing session.

It is interesting to speculate on what could account for the pattern of results observed in Table 2 and Figure 1. One possible explanation is as follows. First, suppose that taking a second form of the examination after practice on a first form produces, on average, an increase in scores (i.e., a practice effect). Second, suppose that taking the computer version of the examination produces, on average, a decrease in scores (i.e., a mode-of-administration effect). Third, suppose that the magnitude of these effects is roughly the same. The combination of practice effect and mode-of-administration effect, as just described, could produce a result like that shown in Figure 1. For the computer-first group, the average score on the first examination (i.e., the computer version) is suppressed due to the mode-of-administration effect. In contrast, the average score on the second examination (i.e., the paper version) is higher, since it increases in two ways: (1) removal of the mode-of-administration effect, and (2) addition of the practice effect. For the paper-first group, the average score for the first examination (i.e., the paper version) is unaffected by either practice or mode-of-administration effects. However, the average score on the second examination (i.e., the computer version) is *increased* by the practice effect, but *decreased*

2. Assuming, of course, that the initial scalings of the paper versions of the examination forms used in this study were correctly carried out.

Table 1. Mathematics Examination Results, Both Versions (Study 1)

Version	N	Mean	Standard Deviation
Computer	94	517.9	80.1
Paper	94	526.9	71.6

by the mode-of-administration effect, resulting in little net change.

Although this explanation is plausible and consistent with the data, alternative explanations, also consistent with the data, are possible. For example, it may be that computer delivery has a negative effect on test scores only for test takers not already familiar with the kinds of questions on the examination. If so, one would expect the average score on the computer version for the computer-first group to be lowered, resulting in a significant difference in scores between computer and paper versions. In contrast, the paper-first group attempts the computer version after having had the opportunity to become familiar with a complete form of the examination. Therefore no reduction in scores is observed for this group on the computer version of the examination.

Table 3 gives Section 1 and Section 2 scaled score means, standard deviations, and the mean difference between the paper and computer versions separately for the computer-first and paper-first groups. For the computer-first group, the mean paper-version Section 1 score was 3.5 points higher than the mean computer-version score. This difference is statistically significant ($t = 5.03$, $df = 46$, $p < .01$). For the paper-first group, the mean paper-version score was 0.6 points higher. This difference is not statistically significant ($t = .83$, $df = 46$). For the computer-first group, the mean paper-version Section 2 score was 1.6 points higher than the mean computer-version score. This difference is statistically significant ($t = 2.12$, $df = 46$, $p < .05$). For the paper-first group, the mean paper-version

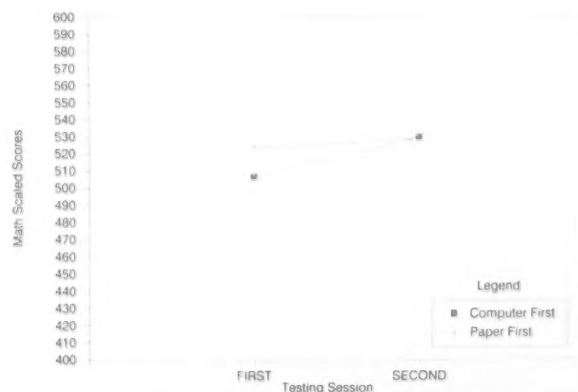


Figure 1. Mathematics total examination scaled scores (Study 1).

Table 2. Mathematics Examination Results, Both Versions, by Order of Administration (Study 1)

Group	N	Computer		Paper		Mean Difference*
		Mean	Standard Deviation	Mean	Standard Deviation	
Computer-first	47	506.6	87.9	529.6	80.3	23.0
Paper-first	47	529.1	70.7	524.1	62.5	-5.0

*Paper-and-pencil score minus computer score.

Table 3. Mathematics Examination Results by Section for Both Versions and Each Group (Study 1)

		Computer		Paper		Mean Difference*
		Mean	Standard Deviation	Mean	Standard Deviation	
Section 1	Computer-first	52.4	8.1	55.9	8.3	3.5
	Paper-first	53.8	7.9	54.4	7.3	0.6
Section 2	Computer-first	50.4	9.4	52.0	9.4	1.6
	Paper-first	53.1	7.5	51.9	7.0	-1.3

*Paper-and-pencil score minus computer score.

score was 1.3 points lower. This difference is not statistically significant ($t = -1.61$, $df = 46$).

Figures 2 and 3 provide plots of the mean Section 1 and Section 2 scores by testing session. The pattern of results for Section 1 (Figure 2) is similar to that observed for the total examination. The computer-first group showed a substantial increase in scores from the first to the second testing. In contrast, the paper-first group showed little net change in scores. The results for the Section 2 scores (Figure 3) are quite different. Both the computer-first and the paper-first groups showed similar increases in scores, and the difference in performance between the two groups was about the same. This kind of pattern is essentially what one would expect if only practice effects, and no mode-of-administration effects, were present. The difference in results for Sections 1 and 2 has two interesting implications regarding possible explanations for the observed pattern of scores. First, it appears that a mode-of-administration effect

is present for Section 1 only. Second, the results for Section 2 suggest that a practice effect is observed for both groups. This casts some doubt on the second explanation for the total examination results described above.

Table 4 provides correlations, coefficient α reliability estimates, and selected estimated true-score correlations for total examination scores and each subscore for both the paper and computer versions. Correlations are given below the main diagonal, reliability coefficients are given in parentheses in the main diagonal, and estimated true-score correlations are given above the main diagonal.

It is evident from Table 4 that, despite possible mode-of-administration and practice effects, the within-group rank orders of scores were quite similar for the computer and paper versions of the examination. The correlation between total examination scores on the paper and computer versions was .90 and .83 for the computer-first and paper-first groups, respectively. The corresponding estimated

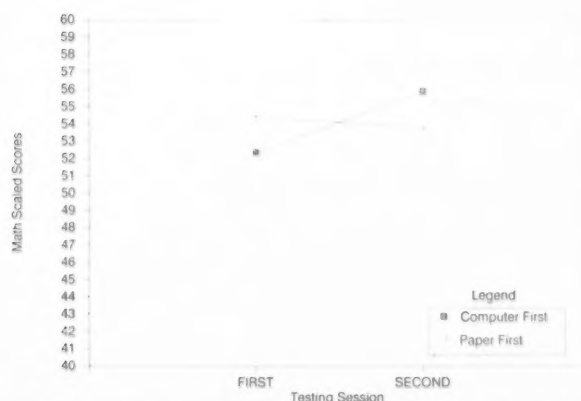
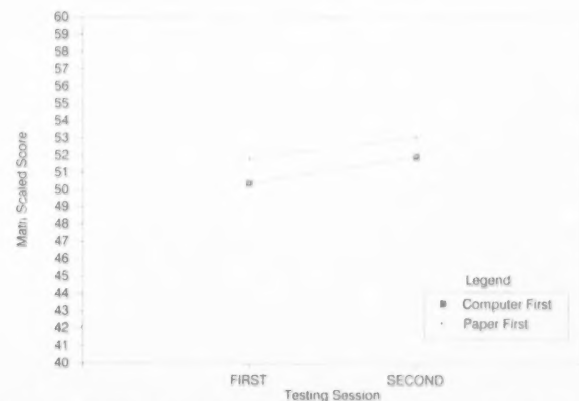
**Figure 2. Mathematics Section 1 scaled scores (Study 1).****Figure 3. Mathematics Section 2 scaled scores (Study 1).**

Table 4. Score Correlations and Reliability Estimates for Mathematics Examination Scores (Study 1)

	Computer-First Group (N = 47)					
	(1)	(2)	(3)	(4)	(5)	(6)
(1) Section 1 paper score	(.85)			1.00*		
(2) Section 2 paper score	.84	(.87)			.99	
(3) Total paper score	.95	.97	(.97)			.98
(4) Section 1 computer score	.87	.86	.88	(.81)		
(5) Section 2 computer score	.76	.86	.85	.85	(.87)	
(6) Total computer score	.82	.90	.90	.95	.97	(.92)

	Paper-First Group (N = 47)					
	(1)	(2)	(3)	(4)	(5)	(6)
(1) Section 1 paper score	(.81)			.97		
(2) Section 2 paper score	.80	(.74)			.96	
(3) Total paper score	.95	.95	(.87)			.95
(4) Section 1 computer score	.77	.73	.79	(.79)		
(5) Section 2 computer score	.72	.73	.76	.74	(.79)	
(6) Total computer score	.80	.78	.83	.92	.94	(.88)

Observed score correlations are given below the diagonal, reliability estimates are in parentheses, and true-score correlations are above the diagonal.

*The estimated true-score correlation slightly exceeded 1.0.

true-score correlations were .98 and .95. Correlations among the subscores were similarly high. For the computer-first group the correlation between the Section 1 subscores was .87, and the correlation between the Section 2 subscores was .86. The corresponding estimated true-score correlations were 1.00 and .99. For the paper-first group the correlation between the Section 1 subscores was .77, and the correlation between the Section 2 subscores was .73. The corresponding estimated true-score correlations were .97 and .96.

Each participant in Study 1 was asked to complete the computer familiarity questionnaire shown in Appendix A. For each question, response options *a* through *e* (or *a* and *b* for question six) were coded from one to five (or from one to two for question six). Recoded responses were then summed to produce a total score for each individual. This total score (CFQ₁) was used in the first study as an index of computer familiarity. The CFQ₁ scores have a possible range of 6 to 27, with higher scores indicating a greater degree of computer familiarity.

Completed questionnaires were obtained from all 94 participants in Study 1 at SWTSU. Table 5 shows the mean and standard deviation of CFQ₁ scores for each of the two study groups. The difference in mean scores between the two groups was not statistically significant.

Table 5. CFQ₁ Scores of Mathematics Examinees (Study 1)

Group	N	Mean	Standard Deviation
Computer-first	47	12.6	4.4
Paper-first	47	12.6	4.2

Table 6 shows the zero-order correlations between CFQ₁ scores and scaled scores on each of the two versions of the CLEP General Examination in Mathematics. The correlations are all relatively small and none are significantly different from zero. One rather surprising entry in Table 6 is that, for the paper-first group, the correlation between CFQ₁ scores and Section 2 computer-version scores is actually negative, which means that students with higher CFQ₁ scores tended to score lower on Section 2 of the computer version of the examination.

Results presented earlier suggested the presence of a mode-of-administration effect on Section 1, but not on Section 2. An additional question of interest was whether the magnitude of such an effect was greater for students who are less familiar with computers. This question was addressed by examining the relationship between CFQ₁ scores and scores on the computer version of the examination, *after controlling for scores on the paper version*. Separate analyses were conducted for Section 1 and Section 2 scores.

For each of the scores a series of four linear multiple-regression models were fit to the data using unweighted least-squares procedures. The first two models were for the regression of computer-version scores on paper-version scores. These models were intended as two alternative baseline models, providing a control for paper-and-pencil scores. Which model to use as a baseline was to be determined from the data.

The models, and requisite notations, were as follows:

Notations:

Y_c = Scores on the computer version

X_p = Scores on the paper version

Table 6. Correlations between CFQ_i Scores and Mathematics Examination Scores (Study 1)

Group	Section 1		Section 2	
	Computer	Paper	Computer	Paper
Computer-first	.18	.15	.06	.17
Paper-first	.03	.21	-.11	.12

Table 7. Estimates of Coefficients and Squared Multiple Correlations for Models I and II for Section 1 and Section 2 Mathematics Examination Scores (Study 1)

	Section 1		Section 2	
	Model I	Model II	Model I	Model II
β_0	4.08**	2.93**	7.29**	5.77**
β_p	.86*	.88*	.83*	.86*
β_e	2.71*	5.32	2.79*	7.05
β_{pe}	—	-.05	—	-.08
R^2	.66	.66	.67	.67

*The parameter estimate is significantly different from 0 at the $\alpha = .01$ level.

**No significance test was done for this parameter estimate.

X_g = Group membership indicator variable, which takes on the value of 0 for the computer-first group and 1 for the paper-first group

Models:

$$(I) \quad Y_c = \beta_0 + \beta_p X_p + \beta_e X_g + \epsilon$$

$$(II) \quad Y_c = \beta_0 + \beta_p X_p + \beta_e X_g + \beta_{pe} X_p X_g + \epsilon$$

Model I is equivalent to fitting separate regression lines for each of the two study groups with a common slope (β_p) but different intercepts (β_0 for the computer-first group and $\beta_0 + \beta_e$ for the paper-first group). Model I has $v_i = n - 4$ degrees of freedom, where n is the total number of observations (in Study 1, $n = 94$). Model II is equivalent to fitting separate regression lines with different intercepts and different slopes (β_p for the computer-first group and $\beta_p + \beta_{pe}$ for the paper-first group). Model II has $v_{II} = n - 5$ degrees of freedom.

Estimates of model coefficients and associated multiple correlations were obtained for each of the two models. As described in Kerlinger and Pedhazur (1973), the test statistic

$$F = \frac{\frac{(R_{II}^2 - R_I^2)}{(v_I - v_{II})}}{\frac{(1 - R_{II}^2)}{v_{II}}}$$

can be referred to an F distribution with $v_I - v_{II}$ degrees of freedom for the numerator and v_{II} for the denominator to determine whether Model II provides a statistically significant improvement in fit over Model I.

Table 7 gives estimates of coefficients, degrees of freedom, and squared multiple correlations (R^2) for Models I and II for both the Section 1 and Section 2 scores.

Model II provided little improvement in fit over Model I for both Section 1 and Section 2 scores, and F tests for both sections indicate that this improvement was not statistically significant at the $\alpha = .05$ level (for Section 1, $F_{1,90} = .456$; for Section 2, $F_{1,90} = .389$). Assuming Model I to be "correct," the slope of the regression line relating computer-version scores to paper-version scores is the same for both groups. However, the estimate of β_e for Section 1 scores indicates that, for a given paper-and-pencil score, the predicted mean score on the computer version of the examination was about 2.7 points higher for the paper-first group than for the computer-first group. The corresponding difference for Section 2 scores was about 2.8 points.

The next step in the analysis was to determine whether the fit of two additional regression models that incorporated CFQ_i scores was significantly better than that provided by Model I. Two additional models were estimated. The models, with required additional notation, were as follows:

Notation:

X_f = Scores on the CFQ_i

Models:

$$(III) \quad Y_c = \beta_0 + \beta_p X_p + \beta_e X_g + \beta_f X_f + \epsilon$$

$$(IV) \quad Y_c = \beta_0 + \beta_p X_p + \beta_e X_g + \beta_f X_f + \beta_{gf} X_g X_f + \epsilon$$

Model III assumes a common partial slope parameter (β_f) relating computer-version scores to CFQ_i scores for both groups. Model IV allows the partial slope parameters

Table 8. Estimates of Regression Coefficients and Squared Multiple Correlations for Models III and IV for Section 1 and Section 2 Mathematics Examination Scores (Study 1)

	Section 1		Section 2	
	Model III	Model IV	Model III	Model IV
β_0	4.57 [†]	2.25 [†]	9.61 [†]	8.52 [†]
β_p	.87**	.87**	.85**	.85**
β_c	2.72**	7.39*	2.80**	5.35
β_e	-.07	.10	-.27*	-.17
β_{cf}	—	-.37	—	-.20
R^2	.66	.67	.68	.69

*The parameter estimate is significantly different from 0 at the $\alpha = .05$ level.

**The parameter estimate is significantly different from 0 at the $\alpha = .01$ level.

[†]No significance test was done for this parameter estimate.

to differ for the two study groups (β_c for the computer-first group and $\beta_c + \beta_{cf}$ for the paper-first group). Models III and IV were both compared to Model I using the F statistic given above to determine whether including CFQ₁ scores resulted in a significant improvement in fit. For those cases in which the addition of CFQ₁ scores did result in a significant improvement in model fit, Model IV was compared to Model III to determine whether a common partial slope parameter could be assumed.

The estimates of coefficients and R^2 that were obtained for both models are shown in Table 8. For Section 1, the addition of CFQ₁ scores did not result in substantial improvement in fit over Model I. The increase in R^2 provided by Model III over Model I was only .002, and the associated F test indicated that this increase was not statistically significant ($F_{1,90} = .311$). The increase in R^2 provided by Model IV over Model I was slightly larger (.009), but was also not statistically significant ($F_{2,89} = 1.37$).

For Section 2, however, CFQ₁ scores did appear to be significantly related to computer-version scores. The increase in R^2 provided by Model III over Model I was not large (.017). However, the associated F test indicated that this increase was statistically significant ($F_{1,90} = 4.857$, $p < .05$). Model IV provided no significant improvement in fit over Model III (increase in $R^2 = .003$, $F_{1,89} = .711$), therefore a common partial slope parameter is assumed for both groups.

There are two puzzling aspects of the results of these regression analyses. One might expect that computer familiarity would be more strongly related to computer-version scores when there was a mode-of-administration effect. However, after controlling for paper-version scores, CFQ₁ scores were found to be significantly related to computer-version scores only for Section 2; earlier analyses (particularly Figures 2 and 3) suggested that mode-of-administration effects were limited to Section 1 scores. Second, one might also expect that, for a given paper-and-pencil score, those who are more familiar with computers would score better on the computer version of the examination. However, the coefficient for CFQ₁ in Model III (i.e., the estimate of β_c) is negative, indicating that, for a given

paper-and-pencil score, the higher the CFQ₁ score, the lower the predicted score on the computer version of the examination. Further discussion of these results appears in the final section of this report.

English Composition Examination

A number of difficulties were encountered in obtaining a matched file of computer- and paper-version examination scores, including item-level data and CFQ₁ responses, for all Study 1 participants at USU. However, we were able to obtain a matched file of total examination computer- and paper-version scores for 116 students (65 in the computer-first group and 51 in the paper-first group) who agreed to participate in Study 1. The scaled score mean and standard deviation for the computer and paper versions of the examination are given in Table 9.

The mean of the computer-version scores was 17.2 points lower than the mean of the paper-version scores. This difference (.271 paper-and-pencil test standard-deviation units) is statistically significant ($t = 3.95$, $df = 115$, $p < .01$). Part of this difference may be due to the greater number of students in the computer-first group. However, if one takes the simple average of the mean score across the two groups for each version, the mean of the computer version is still 16 points lower than the mean of the paper version. This latter mean is not influenced by the different sample sizes across groups. Therefore, as noted before, this difference is probably not explainable in terms of group or order main effects.

Table 10 gives scaled score means, standard deviations, and the mean difference between the paper and

Table 9. English Composition Examination Results, Both Versions (Study 1)

Version	N	Mean	Standard Deviation
Computer	116	469.7	68.3
Paper	116	486.8	63.3

Table 10. English Composition Examination Results, Both Versions, by Order of Administration (Study 1)

Group	N	Computer		Paper		Mean Difference*
		Mean	Standard Deviation	Mean	Standard Deviation	
Computer-first	65	456.8	68.8	489.3	60.4	32.5
Paper-first	51	489.4	64.8	488.9	67.9	-.5

*Paper-and-pencil score minus computer score.

computer versions separately for the computer-first and paper-first groups. For the computer-first group, the mean paper-version score was 32.5 points higher than the mean computer-version score. This difference is statistically significant ($t = 5.77$, $df = 64$, $p < .01$). For the paper-first group, the mean paper-and-pencil score was 0.5 points lower. This difference is not statistically significant ($t = -.09$, $df = 50$).

These results are also presented in Figure 4, which provides a plot of the mean scores given in Table 10 by testing session. The pattern of results observed in Figure 4 is quite similar to that obtained for the Mathematics Examination total scores.

The correlations between scores on the paper and computer versions of the examination were .76 and .82 for the computer-first and the paper-first groups, respectively. Because of difficulties in matching files for the computer and paper examination results, we were unable to obtain the necessary item-level data to generate coefficient α reliability estimates for each of the groups. As a somewhat rough approximation, however, estimates of reliability were obtained using KR-21 (Lord and Novick 1968). For the computer-first group, reliability estimates were .90 for the computer version and .84 for the paper version. Using these estimates in the standard correction for attenuation formulas, we determined that an estimate of the true-score correlations between computer-version and paper-version scores

was .88. For the paper-first group, the estimated reliability was .88 for both versions. Using these estimates, we determined that the estimated true-score correlation between computer-version and paper-version scores was .93.

The values of both estimated true-score correlations for the English Composition Examination, though high, are somewhat lower than those obtained for the Mathematics Examination. This was particularly evident for the computer-first group. These lower estimated true-score correlations raise some concerns about whether the administration of the English Composition Examination by computer may slightly alter the construct being measured. However, it should be remembered that the reliability estimates are only approximate.

The difficulties in matching files with item-level data also precluded doing analyses pertaining to the relationship between CFQ₁ scores and computer-version scores. However, a full set of analyses was conducted for the English Composition Examination in Study 2.

Summary of Study 1 Results

The results from Study 1 suggest that the computer delivery of both the Mathematics and English Composition Examinations results in a mode-of-administration effect. For the groups that took the computer versions first, the average scores for the computer versions of the examinations were significantly lower than the scores for the paper versions. For the groups that took the paper versions first, no significant differences were found between scores on the paper and computer versions. For the Mathematics Examination (where separate Section 1 and Section 2 subscores are reported), this pattern of results appears to be confined to Section 1. In addition, the within-group correlations between the computer version and paper version of the Mathematics Examination are high, and the estimated true scores approach unity. This suggests that the mode-of-administration effect may not affect the nature of what is being measured. However, preliminary correlational analyses for the English Composition Examination indicate that the mode-of-administration effect may slightly alter the construct being measured, particularly for the computer-first group. For the Mathematics Examination, there was no evidence of a relationship between computer-version scores and computer familiarity (as measured by CFQ₁ total scores) for Section 1.

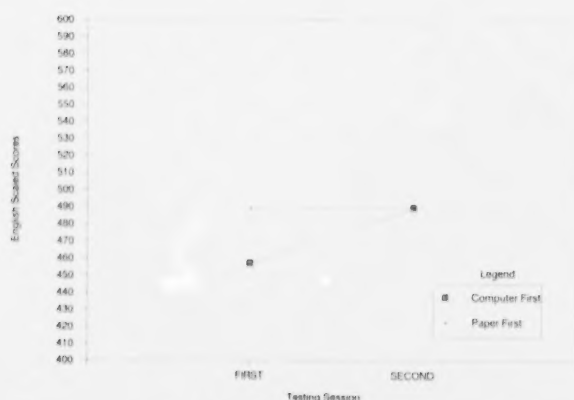


Figure 4. English total examination scaled scores (Study 1).

However, a statistically significant, albeit small, relationship was observed for Section 2 scores. Surprisingly, the direction of this relationship was negative, which implies that for a fixed paper-and-pencil examination, students with higher CFQ₁ scores performed less well on the computer version of the examination.

MODIFICATIONS TO THE COMPUTER VERSIONS FOR STUDY 2

A number of comments were received from the study participants at both institutions regarding difficulties they had with the computer versions of the examinations. In addition, the authors observed a number of additional difficulties with the computer versions that might have contributed to the presence of mode-of-administration effects. Based on this information, several modifications to the computer versions were made.

In Study 1 there was a delay of several seconds between the removal of the current item from the screen and the display of the next item, during which time the computer's clock was running. A number of students expressed concern that they had lost actual exam time because of these delays. An examination of the percentage of participants attempting each item in Study 1 in both the computer and paper versions of the examinations yielded little evidence of speededness for either version. However, based on student feedback, a decision was made to modify the Study 2 computer versions so that the clock was not running during delays between item presentations. Instructions for the examinations were modified to inform students to ignore delays between item presentations, since such delays were not counted as part of their available examination time.

Second, several students commented that they would have liked to mark certain items (or provide provisional answers) for later review (similar to putting a mark in the test booklet next to items one wishes to return to, if time permits). No such feature was available in the computer versions used in Study 1. The Study 1 computer versions did contain a "review screen" feature that listed which items a student had answered, what responses they had given for these items, and which items had been either omitted or not yet attempted. Students could review this listing at any time and could return to any previously attempted item in the section they were currently working on and change their answers. This review screen feature was expanded in Study 2 to allow for marking items. Students were given the capability to identify items they had answered or omitted for later review. The marked items were clearly indicated in the review screen to which students had access throughout a section of the examination.

Third, it was noted that the practice items for the Study 1 computer versions were not limited to the subject area being tested and did not require the students to become familiar with all the features of the computer delivery system.

Thus, for the Study 2 versions, practice items for the Mathematics Examination consisted of sample items from other forms of the CLEP Mathematics Examination, and practice items for the English Composition Examination consisted of sample items from other forms of the CLEP English Composition Examination. In addition, the number of practice items was expanded to include all the item types the students would encounter in the examination.

The Study 1 results regarding the relationship between CFQ₁ scores and computer-version scores were somewhat confusing. After some discussion, it was decided to make some modifications to the questions in the computer familiarity questionnaire and use this revised version in Study 2. Appendix B shows the revised version.

RESULTS OF STUDY 2

Mathematics Examination

A total of 96 students (47 in the computer-first group and 49 in the paper-first group) agreed to participate. The scaled score mean and standard deviation for the computer and paper versions of the examination are given in Table 11. Despite the changes to the computer version, the mean computer-version score was still 6.4 points (.095 paper-and-pencil test standard-deviation units) lower than the mean paper-version score. However, the Study 2 difference is slightly smaller and is not statistically significant ($t = 1.81$, $df = 95$) at the $\alpha = .05$ level.

Table 12 gives scaled score means, standard deviations, and the mean difference between the paper version and the computer version separately for the computer-first and paper-first groups. For the computer-first group, the mean paper-version score was about 21 points higher than the average computer-version score, which is statistically significant ($t = 4.48$, $df = 46$, $p < .01$). For the paper-first group, the mean paper-version score was about seven points lower. This difference is not statistically significant ($t = -1.16$, $df = 48$). Figure 5 provides a plot of the mean scores given in Table 12 by testing session.

A pattern of results was obtained that is quite similar, though not identical, to that of Study 1. As in Study 1, the computer-first group's average score on the second examination (i.e., the paper-version) is substantially higher. However, the paper-first group's average score for the second examination (i.e., the computer version) is also higher, but to a much lesser degree. The increase in scores from the first

Table 11. Mathematics Examination Results, Both Versions (Study 2)

Version	N	Mean	Standard Deviation
Computer	96	513.4	68.8
Paper	96	519.8	67.4

Table 12. Mathematics Examination Results, Both Versions, by Order of Administration (Study 2)

Group	N	Computer		Paper		Mean Difference*
		Mean	Standard Deviation	Mean	Standard Deviation	
Computer-first	47	495.1	64.4	515.8	71.5	20.7
Paper-first	49	530.9	68.5	523.6	63.7	-7.3

*Paper-and-pencil score minus computer score.

Table 13. Mathematics Examination Results by Section for Both Versions and Each Group (Study 2)

Group		Computer		Paper		Mean Difference*
		Mean	Standard Deviation	Mean	Standard Deviation	
Section 1	Computer-first	51.0	6.9	54.1	7.5	3.0
	Paper-first	55.6	7.3	55.5	7.1	-.1
Section 2	Computer-first	49.3	7.4	50.6	8.5	1.3
	Paper-first	52.5	7.6	51.1	7.6	-1.4

*Paper-and-pencil score minus computer score.

testing session to the second for the paper-first group is only about 35 percent as large as the corresponding increase for the computer-first group. This pattern of results would again appear to be consistent with the explanation advanced earlier regarding the operation of mode-of-administration and practice effects.

Table 13 gives Section 1 and Section 2 scaled score means, standard deviations, and the mean difference between the paper and computer versions separately for the computer-first and paper-first groups. For the computer-first group, the mean paper-version Section 1 score was three points higher than the mean computer-version score. This difference is statistically significant ($t = 3.75$, $df = 46$, $p < .01$). For the paper-first group, the mean paper-version score was 0.10 points lower, which is not statistically significant ($t = -.19$, $df = 48$). For the computer-first group,

the mean paper-version Section 2 score was 1.3 points higher than the mean computer-version score. This difference is not statistically significant ($t = 1.95$, $df = 46$) at the $\alpha = .05$ level. For the paper-first group, the mean paper-version score was 1.4 points lower. This difference is statistically significant ($t = -2.19$, $df = 48$, $p < .05$).

Figures 6 and 7 provide plots of the mean Section 1 and Section 2 scores by testing session. As was the case for Study 1, the pattern of results for Section 1 (Figure 6) is similar to that observed for the total examination. The computer-first group showed a substantial increase in scores from the first to the second testing. In contrast, the paper-first group showed little net change in scores. The results for the Section 2 scores (Figure 7) are again different, with both the computer-first and paper-first groups showing similar increases in scores from the first to second testing ses-

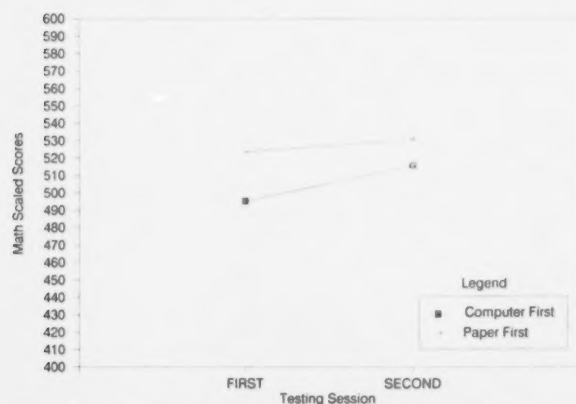


Figure 5. Mathematics total examination scaled scores (Study 2).

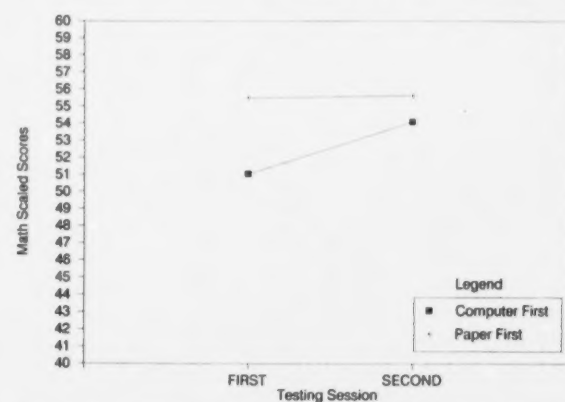


Figure 6. Mathematics Section 1 scaled scores (Study 2).

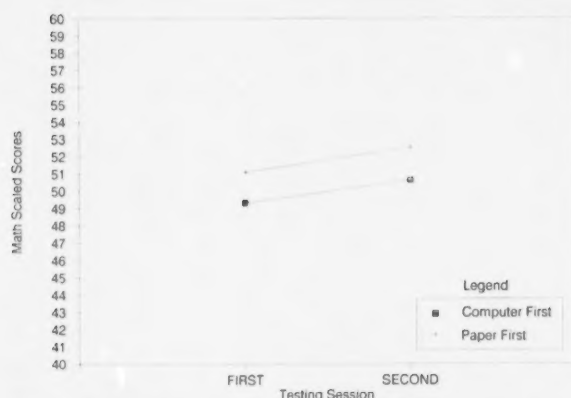


Figure 7. Mathematics Section 2 scaled scores (Study 2).

sions. As noted earlier, this kind of pattern is essentially what one would expect if only practice effects, and no mode-of-administration effects, were present.

Table 14 provides correlations, internal consistency reliability coefficients, and selected estimated true-score correlations for total examination scores and each subscore for both the paper and computer versions of the examination. For the computer-first group, the estimated true-score correlation between computer-version and paper-version Section 1 scores was .94. This is slightly lower than the correlations observed for the Mathematics Examination scores in Study 1. However, correlations for the remaining scores approach unity.

Completed computer familiarity questionnaires were obtained for only 85 of the 96 Study 2 participants: 41 in the computer-first group and 44 in the paper-first group. No statistically significant differences between individuals with

completed CFQs and those without were found for either computer-version or paper-version scores.

As in Study 1, response options *a* through *e* were re-coded as 1 through 5, and scores on the first seven items were summed to produce a total CFQ score (CFQ₂) for each individual. Item 8 was not included in the total score, since the correct ordering of the options was not readily apparent. Table 15 shows the mean and standard deviation of CFQ₂ scores for each of the two study groups.

The difference in mean scores between the two groups was not statistically significant. Table 16 shows the correlation between CFQ₂ scores and scaled scores on each of the two versions of the Mathematics Examination. The correlations are all positive and, in general, considerably larger than those observed in Study 1. Only the correlations of CFQ₂ with Section 1 scores for the paper-first group did not differ significantly from zero. However, there does not appear to be any tendency for CFQ₂ scores to be more highly correlated with computer-version scores than with paper-version scores. In fact, the largest correlation in Table 16 is that between CFQ₂ scores and Section 2 paper-version scores for the computer-first group.

Analyses identical to those in Study 1, examining the relationship between CFQ scores and computer-version scores, were conducted using CFQ₂ scores. Table 17 gives estimates of coefficients, degrees of freedom, and squared multiple correlations for Models I and II for both the Section 1 and Section 2 scores.

As was the case in Study 1, Model II provided little improvement in fit over Model I for both Section 1 and Section 2 scores. In both cases, the increment in *R*² provided by Model II was small (.012 for Section 1 scores and .010 for Section 2 scores). *F* tests (with 1 and 81 degrees of free-

Table 14. Score Correlations and Reliability Estimates for Mathematics Examination Scores (Study 2)

	Computer-First Group (N = 47)					
	(1)	(2)	(3)	(4)	(5)	(6)
(1) Section 1 paper score	(.82)			.94*		
(2) Section 2 paper score	.74	(.84)			1.00*	
(3) Total paper score	.92	.95	(.90)			1.00*
(4) Section 1 computer score	.70	.66	.73	(.69)		
(5) Section 2 computer score	.70	.85	.84	.57	(.81)	
(6) Total computer score	.79	.87	.89	.84	.92	(.85)

	Paper-First Group (N = 49)					
	(1)	(2)	(3)	(4)	(5)	(6)
(1) Section 1 paper score	(.79)			1.00*		
(2) Section 2 paper score	.69	(.79)			1.00*	
(3) Total paper score	.91	.93	(.87)			1.00*
(4) Section 1 computer score	.78	.68	.79	(.74)		
(5) Section 2 computer score	.70	.82	.83	.63	(.82)	
(6) Total computer score	.81	.84	.90	.87	.93	(.87)

Observed score correlations are given below the diagonal, reliability estimates are in parentheses, and true-score correlations are above the diagonal.

*The estimated true-score correlation slightly exceeded 1.0.

Table 15. CFQ₂ Scores of Mathematics Examinees (Study 2)

Group	N	Mean	Standard Deviation
Computer-first	41	16.3	5.0
Paper-first	44	14.6	4.9

dom) for both sections indicate that this small improvement was not statistically significant at the $\alpha = .05$ level (for Section 1, $F = 2.667$; for Section 2, $F = 1.765$). Assuming Model I to be "correct," the slope of the regression line relating computer-version scores to paper-version scores is the same for both groups. However, the estimate of β_x for Section 1 scores indicates that, for a given paper-and-pencil score, the predicted mean score on the computer version of the examination is about 3.6 points higher for the paper-first group than for the computer-first group. The corresponding difference for Section 2 scores is about 2.5 points.

The estimates of coefficients and R^2 obtained for Models III and IV are given in Table 18. For both Section 1 and

Section 2 scores, the addition of CFQ₂ scores did not result in substantial improvement in fit over Model I. The increase in R^2 provided by Model III over Model I was only .006 for Section 1 and .011 for Section 2. The associated F tests indicated that neither increase was statistically significant ($F_{1,81} = .029$ for Section 1, $F_{1,81} = .143$ for Section 2). The increases in R^2 provided by Model IV over Model I were slightly larger, but were also not statistically significant. Therefore, unlike in Study 1, there was no evidence of a relationship between computer familiarity, as measured by CFQ₂ scores, and computer-version scores after controlling for scores on the paper version.

English Composition Examination

A total of 115 students (58 in the computer-first group and 57 in the paper-first group) agreed to participate in Study 2. The scaled score mean and standard deviation for the computer and paper versions of the examination are given in Table 19. Unlike Study 1, there was almost no difference in

Table 16. Correlations between CFQ₂ Scores and Mathematics Examination Scores (Study 2)

Group	Section 1		Section 2	
	Computer	Paper	Computer	Paper
Computer-first	.42	.40	.36	.50
Paper-first	.14	.29	.38	.35

Table 17. Estimates of Coefficients and Squared Multiple Correlations for Models I and II for Section 1 and Section 2 Mathematics Examination Scores (Study 2)

	Section 1		Section 2	
	Model I	Model II	Model I	Model II
β_0	12.31**	16.91**	11.57**	14.55**
β_p	.72*	.63*	.75*	.70*
β_c	3.59*	-8.86	2.54*	-6.64
β_{pc}	—	.02	—	.02
R^2	.61	.62	.69	.70

*The parameter estimate is significantly different from 0 at the $\alpha = .01$ level.

**No significance test was done for this parameter estimate.

Table 18. Estimates of Regression Coefficients and Squared Multiple Correlations for Models III and IV for Section 1 and Section 2 Mathematics Examination Scores (Study 2)

	Section 1		Section 2	
	Model III	Model IV	Model III	Model IV
β_0	12.28†	10.71†	11.53†	13.44†
β_p	.71**	.71**	.74**	.75**
β_c	3.63**	6.88*	2.62**	-1.79
β_{pc}	.02	.14	.04*	-.12
β_{cf}	—	-.23	—	.31
R^2	.61	.61	.69	.70

*The parameter estimate is significantly different from 0 at the $\alpha = .05$ level.

**The parameter estimate is significantly different from 0 at the $\alpha = .01$ level.

†No significance test was done for this parameter estimate.

Table 19. English Composition Examination Results, Both Versions (Study 2)

Version	N	Mean	Standard Deviation
Computer	115	485.4	79.7
Paper	115	485.8	80.7

score means for the computer version and the paper version. The mean paper-version score was only 0.4 points higher than the average computer-version score, and the difference was clearly not statistically significant ($t = .08$, $df = 114$).

Table 20 gives scaled score means, standard deviations, and the mean difference between the paper and computer versions separately for the computer-first and paper-first groups. For the computer-first group, the mean paper-version score was about 13 points higher than the mean computer-version score. This difference is not statistically significant ($t = 1.71$, $df = 57$). For the paper-first group, the mean paper-version score was about 12 points lower. This difference is statistically significant ($t = -2.18$, $df = 56$, $p < .05$).

Figure 8 provides a plot of the mean scores given in Table 20 by testing session. Figure 8 looks quite different from the corresponding figure for Study 1 (see Figure 4). In fact, the pattern of results looks similar to what was observed for Section 2 scores on the Mathematics Examination (see Figure 7). Both the computer-first and paper-first groups show similar increases in scores from the first to second testing session. As noted earlier, this kind of pattern is essentially what one would expect if only practice effects, and no mode-of-administration effects, were present.

For Study 2, we were successful in matching item response data for the paper and computer versions of the examination. For the computer-first group, the estimated coefficient α reliabilities for the paper and computer versions were .91 and .89, respectively. The correlation between scores on the paper and computer versions was .78. The estimated true-score correlation was .87. As with Study 1, this correlation is disturbingly low and suggests that, even though the mode-of-administration effect on mean scores may have been eliminated, the computer delivery of the test may still be having an effect on the nature of the construct being measured.

Table 20. English Composition Examination Results, Both Versions, by Order of Administration (Study 2)

Group	N	Computer		Paper		Mean Difference*
		Mean	Standard Deviation	Mean	Standard Deviation	
Computer-first	58	481.2	83.9	493.9	86.4	12.7
Paper-first	57	489.8	75.7	477.6	74.4	-12.2

*Paper-and-pencil score minus computer score.

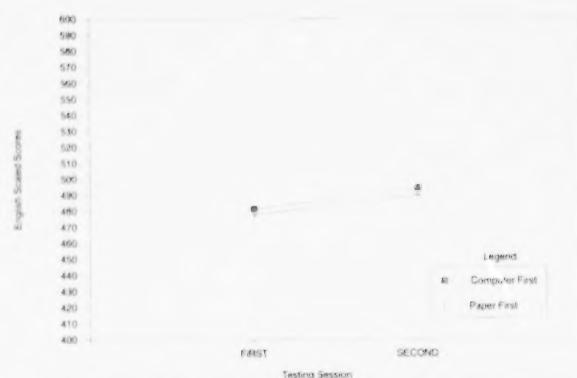


Figure 8. English total examination scaled scores (Study 2).

For the paper-first group, the estimated coefficient α reliability was .88 for both versions of the examination. The correlation between scores on the paper and computer versions was .843. The estimated true-score correlation was .96, which is much more in line with the results for the Mathematics Examination.

CFQ₂ scores were available for 103 of the 115 participants in Study 2 (53 from the computer-first group and 50 from the paper-first group). As was the case for the Mathematics Examination, no statistically significant differences between individuals with completed CFQs and those without were found for either computer-version or paper-version scores.

Table 21 shows the mean and standard deviation of CFQ₂ scores for the two study groups. The difference in mean scores is small and not statistically significant. Table 22 shows the correlation between CFQ₂ scores and scaled scores on each of the two versions of the English Composition Examination. As was the case for the Mathematics Examination in Study 1, all zero-order correlations were small and were not significantly different from zero.

Table 23 gives estimates of coefficients, degrees of freedom, and squared multiple correlations for Models I and II. As was the case in both studies for the Mathematics Examination, Model II provided little improvement in fit over Model I. The increment in R^2 provided by Model II was small (.008), and the associated F test (with 1 and 99 degrees of freedom) indicates that this small improvement was not statistically significant at the $\alpha = .05$ level ($F =$

Table 21. CFQ₂ Scores of English Composition Examinees (Study 2)

Group	N	Mean	Standard Deviation
Computer-first	53	14.9	4.6
Paper-first	50	15.7	4.5

Table 23. Estimates of Coefficients and Squared Multiple Correlations for Models I and II for English Composition Examination Scores (Study 2)

	Model I	Model II
β_0	82.17 [†]	114.34 [†]
β_p	.81**	.59**
β_c	18.18*	-54.75
β_{pc}	—	.15
R^2	.66	.67

**t*-test indicates that the parameter estimate is significantly different from 0 at the $\alpha = .05$ level.

***t*-test indicates that the parameter estimate is significantly different from 0 at the $\alpha = .01$ level.

[†]No significance test was done for this parameter estimate.

1.67). Assuming Model I to be "correct," the slope of the regression line relating computer-version scores to paper-version scores is the same for both groups. However, the estimate of β_c for total examination scores indicates that, for a given paper-version score, the predicted mean score on the computer version of the examination is about 18.2 points higher for the paper-first group than for the computer-first group.

The estimates of coefficients and R^2 obtained for Models III and IV are given in Table 24. The addition of CFQ₂ scores again did not result in substantial improvement in fit over Model I. The increase in R^2 provided by Model III over Model I was only .006. The associated *F* tests indicated that this increase was not statistically significant ($F_{1,99} = 1.0614$). The increase in R^2 provided by Model IV over Model I was larger (.015), but was also not statistically significant ($F = 2.350$). As was the case for the Mathematics Examination scores in Study 2, there was little evidence of a relationship between computer familiarity, as measured by CFQ₂ scores, and computer-version scores after controlling for scores on the paper version.

Summary of Study 2 Results

It appears that, for the Mathematics Examination, the changes made to the computer version did not eliminate the mode-of-administration effect on mean scores for Section I. In contrast, for the English Composition Examination, average scores from the two versions were almost identical, suggesting that the mode-of-administration effect on mean scores had been removed. However, a troubling finding was that the estimated true-score correlation between computer versions and paper versions was somewhat lower than might

Table 22. Correlations between CFQ₂ Scores and English Composition Examination Scores (Study 2)

Group	Computer	Paper
Computer-first	.12	-.06
Paper-first	.05	.11

Table 24. Estimates of Regression Coefficients and Squared Multiple Correlations for Models III and IV for English Composition Examination Scores (Study 2)

	Model III	Model IV
β_0	63.31 [†]	33.85 [†]
β_p	.81**	.82**
β_c	17.20	71.95*
β_{pc}	1.31	6.58*
β_{cp}	—	-3.57
R^2	.67	.68

**t*-test indicates that the parameter estimate is significantly different from 0 at the $\alpha = .05$ level.

***t*-test indicates that the parameter estimate is significantly different from 0 at the $\alpha = .01$ level.

[†]No significance test was done for this parameter estimate.

be expected for the computer-first group. For both examinations there was little evidence of a relationship between computer familiarity (as measured by CFQ₂ total scores) and scores on the computer versions of the examinations after controlling for paper-version scores.

DISCUSSION AND CONCLUSIONS

The results of both studies reported here suggest that, despite efforts to design computer versions of the CLEP Mathematics and English Composition General Examinations that were administratively similar to paper-and-pencil testing (i.e., allowed item review and answer changing and were comparably timed), mode-of-administration effects were found. It should be emphasized that mode-of-administration effects were found for two examinations with relatively simple item types. In addition, the populations of examinees taking the tests at the two institutions were somewhat different. Besides the differences in the demographic makeup of students attending state-funded schools in Texas and Utah, the students at SWTSU were primarily sophomores and juniors taking a mandatory "high-stakes" examination (upper-level status was contingent on passing the examination). In contrast, a large number of students at USU were freshmen, and the examinations were offered on a voluntary basis. The results of these studies underscore the need to determine empirically (rather than just to assume) the equivalence of computer and paper versions of an examination.

The Study 2 results suggest that the modifications made to the computer delivery system may have reduced the mode-of-administration effect on the average scores for the

English Composition Examination. However, the estimated true-score correlations between scores on the paper and computer versions are slightly lower than would be expected (at least for the computer-first group), suggesting that the mode of test delivery may still be having an effect on the nature of the construct being measured.

Mode-of-administration effects on average scores still appear to be present in Section 1 of the Mathematics Examination. However, there is some indication that the magnitude of this Section 1 mode-of-administration effect may have been reduced slightly in Study 2. For the computer-first group, the increase in scores from the first to second testing session is about 0.5 points less for the Study 2 data than the Study 1 data. This difference could reflect a reduction in the size of the mode-of-administration effect. For the paper-first group, scores *decreased* about 0.6 points from the first testing session to the second in Study 1, but *increased* about 0.1 points in Study 2. This is a net change of about 0.7 points and would be consistent with a reduction in the magnitude of the mode-of-administration effect.

Of course, it should be noted that an alternative explanation is possible. Mode-of-administration effects may have stayed constant and practice effects increased, which is also consistent with this pattern of results. However, there is no *a priori* reason to expect practice effects to change from one study to the next, and changes to the Study 2 computer versions were made with the hope of reducing the mode-of-administration effect. In either case, the evidence for reduction in mode-of-administration effect (or increase in practice effect) is not strong, given the small magnitude of the change from Study 1 to Study 2 and the relatively modest sample sizes in both studies.

Taken as a whole, there does not appear to be compelling evidence for a relationship between computer familiarity and computer-version scores after controlling for paper-version scores. Although a statistically significant partial association was obtained in Study 1 between CFQ₁ scores and computer-version scores on Section 2 of the Mathematics Examination, the strength of the association, as measured by the increase in R^2 , was fairly small. In addition, similar results were not observed for Section 1 scores in Study 1, nor for any of the Mathematics or English Composition Examination scores in Study 2.

There are at least two other reasons why the one significant partial association that was obtained should be viewed cautiously. First, the direction of the association is negative, meaning that, for a fixed paper-and-pencil score, the students reporting greater degrees of computer familiarity performed worse on the computer version of the examination. It seems, at least to these authors, that if an association did exist, that association would be positive. That is, for a group of students of equivalent mathematics ability (as measured by the paper-and-pencil version of the examination), those students with greater computer familiarity would be aided by their experience and score somewhat higher on the computer version of the examination. Of course it is possible to generate explanations of why the association might

be negative. For example, students with a lot of experience operating word processors might have strong associations regarding the functions of certain keys. If these keys performed different functions in the computer version of the examination, these previous associations might have interfered with learning the computer test delivery system and somehow impeded performance.

The significant partial association obtained in Study 1 is anomalous in a second way. Analyses of mean scores suggest that an average mode-of-administration effect existed for Section 1 of the Mathematics Examination, but not for Section 2. It seems to the authors of this report that one would be more likely to observe a significant partial association between CFQ₁ scores and Section 1 scores when mode-of-administration effects are prominent. However, this did not turn out to be the case. The single significant association between CFQ₁ scores and computer-version scores was on Section 2.

Despite these somewhat anomalous features, the one instance of association between computer-version scores and computer familiarity should not be completely dismissed. Because the sample sizes were small, the studies described in this report do not have a lot of statistical power to detect associations that are small in magnitude. In addition, the computer familiarity questionnaires used in the studies were relatively short and were not field-tested or validated before their use in the current research. Perhaps scores obtained from longer and more carefully constructed questionnaires would exhibit a greater degree of association with computer-version scores.

Based on the results of the current studies, particularly the moderately encouraging results of Study 2, CLEP is continuing to modify the examinee-computer interface in order to continue to reduce and, it is hoped, eliminate mode-of-administration effects. Additional modifications to the computer delivery system are in progress. The modifications include conversion to a VGA monitor standard to improve the quality of display; conversion to a mouse response device, and associated system of icons, to simplify the examinee-computer interface; and an expanded tutorial system for introducing examinees to the features of the computer delivery system. When these are completed, further studies will be conducted to evaluate the degree of comparability of scores on the paper and computer versions of the examinations.

REFERENCES

- American Psychological Association. 1986. *Guidelines for Computer-Based Tests and Interpretations*. Washington, D.C.: American Psychological Association.
- Angoff, W. H. 1984. *Scales, Norms, and Equivalent Scores*. Princeton, N.J.: Educational Testing Service.
- Bunderson, V. C., D. K. Inouye, and J. B. Olsen. 1989. "The Four Generations of Computerized Educational Measurement." In *Educational Measurement* (3d edition), edited by R. L. Linn. New York: Macmillan.

- College Board. 1978. *CLEP Test Information Guide, General Examination in Mathematics*. New York: College Entrance Examination Board.
- College Board. 1986. *CLEP Test Information Guide, General Examination in English Composition*. New York: College Entrance Examination Board.
- Kerlinger, F. N., and E. J. Pedhazur. 1973. *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart & Winston.
- Lee, J. 1986. "The Effects of Past Computer Experience on Computerized Aptitude Test Performance." *Educational and Psychological Measurement* 46:727-734.

- Lord, F. M., and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.
- Mazzeo, J., and A. L. Harvey. 1988. *The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests: A Review of the Literature*. (College Board Report No. 88-8, ETS RR No. 88-21). New York: College Entrance Examination Board.
- Ward, W. C. 1984. "Using Microcomputers to Administer Tests." *Educational Measurement Issues and Practice*. 3:16-20.

APPENDIX A. COMPUTER FAMILIARITY QUESTIONNAIRE (STUDY 1)

Name: _____
 Last First M.I.

Social Security Number: _____

1. How many computer courses have you completed?
 - (a) 0
 - (b) 1
 - (c) 2
 - (d) 3
 - (e) more than 3
2. In the past year, how many letters or reports have you written using a word processor?
 - (a) 0
 - (b) 1
 - (c) 2
 - (d) 3
 - (e) more than 3
3. In the past year, how often have you played computerized games, such as Pac-Man?
 - (a) never
 - (b) approximately 1-11 times
 - (c) occasionally but not regularly, approximately 11 times to twice a month
 - (d) regularly, approximately twice a month to every week
 - (e) at least once a week
4. In the past year, you have completed how many academic courses and/or job training courses (other than computer courses) in which you typed data (numerical or prose or symbolic) into a computer?
 - (a) 0
 - (b) 1
 - (c) 2
 - (d) 3
 - (e) more than 3
5. In any of your jobs in the past year, how often have you typed data into a computer?
 - (a) never
 - (b) approximately 1-11 times
 - (c) occasionally but not regularly, approximately 11 times to twice a month
 - (d) regularly, approximately twice a month to every week
 - (e) at least once a week
6. Have you ever taken a computer-administered exam before?
 - (a) no
 - (b) yes

APPENDIX B. COMPUTER FAMILIARITY QUESTIONNAIRE (STUDY 2)

Name: _____

Social Security Number: _____

1. How many computer courses have you completed?

- (a) 0
- (b) 1
- (c) 2
- (d) 3
- (e) more than 3

2. How many computer languages (e.g., BASIC, PASCAL, FORTRAN) have you written programs in?

- (a) 0
- (b) 1
- (c) 2
- (d) 3
- (e) more than 3

3. In the past year, how many letters or reports have you written using a word processor?

- (a) 0
- (b) 1 to 3
- (c) 4 to 6
- (d) 7 to 12
- (e) more than 12

4. How many word-processing packages (e.g., *WordPerfect*, *Displaywrite*, *PCWrite*, *Microsoft Word*) have you used to produce documents (such as letters or reports)?

- (a) 0
- (b) 1
- (c) 2
- (d) 3
- (e) more than 3

5. In the past year, how often have you used other commercially available software packages such as spread-

sheets (e.g., Lotus) or statistical packages (e.g., SPSS, BMDP, SAS)?

- (a) 0
- (b) 1 to 3
- (c) 4 to 6
- (d) 7 to 12
- (e) more than 12

6. In the past year, how often have you been required to simply type data (either numerical or symbolic) into a computer?

- (1) 0
- (b) 1 to 3
- (c) 4 to 6
- (d) 7 to 12
- (e) more than 12

7. Prior to the current test, how many computer-administered exams have you taken?

- (a) 0
- (b) 1
- (c) 2
- (d) 3
- (e) more than 3

8. How would you best characterize how you have acquired your experience with computers?

- (a) completely self-taught
- (b) primarily self-taught, with some formal coursework
- (c) primarily through formal coursework, with some self-teaching
- (d) completely through formal coursework
- (e) I have no experience with computers